机器学习中的数学第 5 课：线性代数初步

管枫

七月在线

June, 2016

主要内容

- 线性空间与基
  - 举例说明线性空间的基本概念
- 线性映射与矩阵
  - 什么是矩阵?
  - 矩阵作为线性映射的代数表达方式
  - 线性方程的几何意义
- 线性回归
  - 线性回归作为方程求解问题
  - 线性回归作为几何逼近问题
  - 最小二乘法
- 相似不变量

记号

- 本节课常用数学记号

V, W	向量空间
v, w	向量
$\mathbb{R}^n$,$\mathbb{R}^m$	实坐标空间
α,β	V 和 W 的基
$T : V \to W$	向量空间 V 到 W 的线性映射
$A_{\alpha,\beta}(T)$	线性映射 T 在 α 和 β 这两组基下的矩阵
$G(v_1, v_2)$	内积空间 V 上的内积
H_α	G 在基 α 下的矩阵形式

线性空间与基

实系数线性空间是一个由向量组成的集合, 向量之间可以做加减法, 向量与实数之间可以做乘法, 而且这些加，减，乘运算要求满足常见的交换律和结合律. 我们也可以类似地定义其他系数的线性空间.

Example (线性空间)

有原点的平面。

- 如果平面有一个原点 O, 那么平面上任何一个点 P, 都对应着一个向量 $\overrightarrow{OP}$。
- 这些向量以及他们的运算结构放在一起，就组成一个向量空间。

- 原点 O 在空间中引入了线性结构。(向量之间的加法，以及向量与实数的乘法)

线性空间与基

基是线性空间里的一组向量，使得任何一个向量都可以唯一的表示成这组基的线性组合.

Example (坐标空间)

有原点的平面，加上一组基 $\{\vec{X}, \vec{Y}\}$。

- 任何一个向量 $\overrightarrow{OP}$, 都可以唯一表达成 $\overrightarrow{OP} = a\vec{X} + b\vec{Y}$ 的形式。
- (a, b) 就是 P 点的坐标。
- 基给出了定量描述线性结构的方法——坐标系。

线性空间与基

Example (坐标系的选择)

考虑纽约中城区的地图。街道用数字编号，但不是正南正北。在这个地图上以中心为原点，如何选取基?

- 基的选择取决于要解决的问题。
- 没有十全十美的基，只有适合解决问题的基。

线性空间与基

小结 (线性空间与基)

- 线性空间是一种结构 (加法及乘法运算结构)
- 基使得我们可以用坐标描述线性结构
- 基的选择取决于要研究的问题

线性映射与矩阵

Definition (线性映射)

V 和 W 是两个实线性空间，$T : V \to W$ 如果满足如下条件就是一个线性映射。

$$
\begin{aligned}
&(i) \quad T(v_1 + v_2) = T(v_1) + T(v_2), && \forall v_1, v_2 \in V \\
&(ii) \quad T(\lambda v) = \lambda T(v), && \forall \lambda \in \mathbb{R}, v \in V
\end{aligned}
$$

- 线性映射的本质就是保持线性结构的映射
- 到自身的线性映射 $T : V \to V$ 叫做线性变换

线性映射与矩阵

线性变换的矩阵描述

V, W 分别为 n, m 维的线性空间，
$\alpha = \{\alpha_1, ..., \alpha_n\}, \beta = \{\beta_1, ..., \beta_m\}$ 分别为 V, W 的一组基。
$T : V \to W$ 是一个线性映射。于是 T, α, β 唯一决定一个矩阵 $A_{\alpha,\beta}(T) = [A_{ij}]_{m\times n}$，使得

$$T(\alpha_j) = \sum_{i=1}^{m} A_{ij} * \beta_i, \forall j \in 1, ..., n \tag{1}$$

(1) 等价于

$$T(\alpha_1, ..., \alpha_n) = (\beta_1, ..., \beta_m) \cdot A_{\alpha,\beta}(T) \tag{2}$$

简记为

$$T(\alpha) = \beta \cdot A_{\alpha,\beta}(T) \tag{3}$$

线性映射与矩阵

如果我们选取 V, W 的另外一组基，$\tilde{\alpha} = \alpha \cdot P$, $\tilde{\beta} = \beta \cdot Q$. 那么存在矩阵 $A_{\tilde{\alpha},\tilde{\beta}}(T)$ 使得,

$$T(\tilde{\alpha}) = \tilde{\beta} \cdot A_{\tilde{\alpha},\tilde{\beta}}(T)$$

两边分别代入 $\tilde{\alpha}$ 与 $\tilde{\beta}$ 得到,

$$T(\alpha) \cdot P = T(\alpha \cdot P) = \beta \cdot Q \cdot A_{\tilde{\alpha},\tilde{\beta}}(T)$$

与(3)比较我们得到矩阵变换公式:

$$Q \cdot A_{\tilde{\alpha},\tilde{\beta}}(T) \cdot P^{-1} = A_{\alpha,\beta}(T) \tag{4}$$

小结 (线性映射与矩阵)

- 矩阵是线性映射在特定基下的一种定量描述

$$T(\alpha) = \beta \cdot A_{\alpha,\beta}(T)$$

- 基变换下的矩阵变换公式的推导方法

$$Q \cdot A_{\tilde{\alpha},\tilde{\beta}}(T) \cdot P^{-1} = A_{\alpha,\beta}(T)$$

线性映射与矩阵

例题 (几何变换：拉伸, 反转与旋转)

实数平面 $\mathbb{R}^2$ 到自身的映射：

- 拉伸 $T\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$
- 反转 $T\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$
- 旋转 $T\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \cos(\pi/6) & -\sin(\pi/6) \\ \sin(\pi/6) & \cos(\pi/6) \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

Example (算法问题：如何计算斐波那契数列)

斐波那契数列：$a_1 = 1, a_2 = 1$, 并且满足 $a_{n+2} = a_n + a_{n+1}$.

Solution (1. 简单递归)

我们可以使用定义进行简单递归

```
1 def arrayComputer(n):
      if n <= 2:
3         return 1
      else:
5         prevA = arrayComputer(n-1)
          prevprevA = arrayComputer(n-2)
7         return prevA + prevprevA
```

Solution (2. 建立线性模型)

把 $[a_n, a_{n+1}]^T$ 看成一个向量，于是递归公式变成一个线性模型：

$$\begin{bmatrix} a_{n+1} \\ a_{n+2} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} a_n \\ a_{n+1} \end{bmatrix}$$

于是：

$$\begin{bmatrix} a_n \\ a_{n+1} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}^{n-1} \cdot \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$

线性模型算法的 Python code:

```
def linearComputer(n):
    if n <= 2:
        return 1
    current2A = [1, 1]
    prev2A = copy.deepcopy(current2A)
    for ind in range(n-2):
        current2A[0] = 0*prev2A[0] + 1*prev2A[1]
        current2A[1] = 1*prev2A[0] + 1*prev2A[1]
        prev2A = copy.deepcopy(current2A)
    return prev2A[1]
```

思考题

如何计算类似的数列 $a_1 = 1, a_2 = 1, a_3 = 1$, 并且 $a_{n+3} = a_n + 2a_{n+1} + 3a_{n+2}$?

线性回归

线性回归

- 模型:$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$, 余项 ϵ 服从正态分布 $N(0, \sigma)$
- 数据: 样本 $(Y_i, X_{i1}, \cdots, X_{ik}), 1 \leq i \leq n$
- 目的: 估计参数 $\beta_1, \cdots, \beta_k$

线性回归: 矩阵模型

可以把整个样本用矩阵表示。令 $x_0 = 1$, 原模型变成

$$y = \beta_0 x_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

用矩阵 Y, X, β 定义如下

$$\begin{aligned} Y &= [Y_1, \cdots, Y_n]^T \\ X &= [X_{ij}]_{1 \le i \le n, 0 \le j \le k} \\ \beta &= [\beta_0, \cdots, \beta_k]^T \end{aligned}$$

于是整个模型写成:

$$Y = X \cdot \beta + \epsilon$$

线性回归: 线性方程（代数）

如果我们省略掉最后一个误差项，问题变为解线性方程的问题

$$X \cdot \beta = Y \tag{5}$$

一般来讲，样本个数大于自参数个数。所以方程个数大于这个方程的未知数个数，于是方程通常是没有解，长方形矩阵也一定没有逆矩阵。但是如果 $X^T X$ 是可逆矩阵 (一般是满足的)，那么代数上可以用如下方法求一个近似的解答。

$$X^T X \cdot \beta = X^T Y$$
$$\beta = (X^T X)^{-1} X^T Y$$

所以如若(5)有解，就一定是这个 $\beta = (X^T X)^{-1} X^T Y$. 而如果(5)没有解，这个 β 也是一个合理的估计.

线性回归: 几何逼近（几何）

我们从线性映射的角度重新来审视这个方程

$$X \cdot \beta = Y \tag{6}$$

等式的左边是矩阵 X 的列向量的一个线性组合，所以这个方程的几何意义是希望把 Y 当成 X 列空间中的一个点，然后求出这个点在列向量这组基下的坐标.

——————————但是!!——————————

Y 不见得是 X 的列空间中的点啊，怎么办? 几何学家的想法是找到 Y 在 X 列空间里的投影 Y^*. 然后用 Y^* 在列空间上的坐标来估计 β。换句话说希望找到 β 使得 $X\beta - Y$ 与 X 垂直，也就是说 $X^T \cdot (X\beta - Y) = 0$, 于是乎

$$X^T X\beta = X^T Y$$

$$\beta = (X^T X)^{-1} X^T Y$$

线性回归: 最小二乘（统计）

统计学家采取更加简单直接的想法，线性模型最终是用来做预测的，那么预测的准确程度才是这个模型优良的最终度量. 所以统计学家决定最小化误差项 $Y - X\beta$. 也就是 $(Y - X\beta)^T \cdot (Y - X\beta)$. 这是一个关于 β 的二次型，关于二次型的更加深入内容我们下次来讲解，不过这里我们先用它来解决一下线性回归的最小二乘方法。为了求极值，我们对 β 求导数（梯度）。

$$\nabla (Y - X\beta)^T \cdot (Y - X\beta) = -2X^T Y + 2X^T X\beta$$

极值条件为 $\nabla (Y - X\beta)^T \cdot (Y - X\beta) = 0$, 于是我们得到方程

$$-2X^T Y + 2X^T X = 0$$

再一次我们得到

$$\beta = (X^T X)^{-1} X^T Y$$

线性回归: 极大似然估计（统计）

统计学家也可以采取极大似然估计的方法[1]. 将模型转化为概率分布的参数估计问题.$y(\sum_{j=o}^{k} x_j\beta_j, \sigma)$，利用我们的样本进行极大似然估计. 似然函数为

$$L(\beta) = \prod_{i=1}^{n} exp(\frac{(\sum_{j=0}^{k} x_j\beta_j - y)^2}{2\sigma^2})$$

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} \frac{(\sum_{j=0}^{k} x_j\beta_j - y)^2}{2\sigma^2} \\ &= \frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta) \end{aligned}$$

于是得到了和前面最小二乘法一样的优化函数. 值得注意的是，这种方法更具备一般性，即使不是线性模型也可以使用.

[1]感谢 @Zhui 同学提供的材料: 程序员的数学 2：概率统计，平冈和幸堀玄 (著)，陈筱烟 (译)

矩阵的标准型: 方阵的相似变换

如果 $T: V \to V$ 是一个线性变换, 那么对于 V 的两组基 α 与 $\tilde{\alpha} = \alpha \cdot P$，线性变换 T 的矩阵分别为

$$A_\alpha(T) \text{ and } A_{\tilde{\alpha}}(T) = P^{-1} \cdot A_\alpha(T) \cdot P$$

方阵的相似变换

- 如果两个方阵 A 和 $\tilde{A}$ 满足，$\tilde{A} = P^{-1}AP$. 那么这两个方阵就互为相似矩阵
- 相似矩阵的几何意义是同一个线性变换在不同的基下的表达形式
- 当研究对象是线性变换的时候，我们只关心矩阵在相似变换下不变的几何性质。

矩阵的标准型

相似变换下不变的性质

本节列举一些相似不变量，稍后利用约当标准型来介绍他们的几何意义。

- 行列式 (det)

$$\begin{aligned}\det(P^{-1}AP) &= \det(P^{-1})\det(A)\det(P)\\ &= \det(P^{-1})\det(P)\det(A)\\ &= \det(A)\end{aligned}$$

- 迹 (trace),$\mathrm{tr}(AB) = \mathrm{tr}(BA)$

$$\mathrm{tr}(P^{-1}AP) = \mathrm{tr}(APP^{-1}) = \mathrm{tr}(A \cdot I) = \mathrm{tr}(A)$$

- 秩 (rank)

矩阵的标准型: 相似不变量

相似变换下不变的性质

- 特征值：特征方程 $\det(A-\lambda I)=0$ 的根。
  如果 $\det(A-\lambda I)=0$，那么 $\det(P^{-1}(A-\lambda I)P)=0$, 于是 $\det(P^{-1}AP-\lambda I)=0$

- 特征值是最重要的相似不变量，利用这个相似不变量可以方便的得出上面所有的不变量。

谢谢大家!